VII. Advanced topics

- It is labelled mainly because they may not fit in earlier sessions, thus it can be seen as an open session for discussion.
- Technically this would include
  - Bayesian networking
  - Structural equation modelling
  - methods on dimension reduction
- They may largely be presented as case studies.

Analysis of Pathways

- Statistical Methods
- Example Analyses
- A Case Study
  - Methods
  - Results
  - Summary

Statistical Methods

- Graphic models
  - WinBUGS
  - gR
- Bayesian networks
- Structural equation modelling
  - SPSS/AMOS, LISREL, Mx, EQS, MPlus, SYSTAT/EzPATH, SAS/CALIS, SAS/SYSLIN
  - R/sem, R/systemfit,

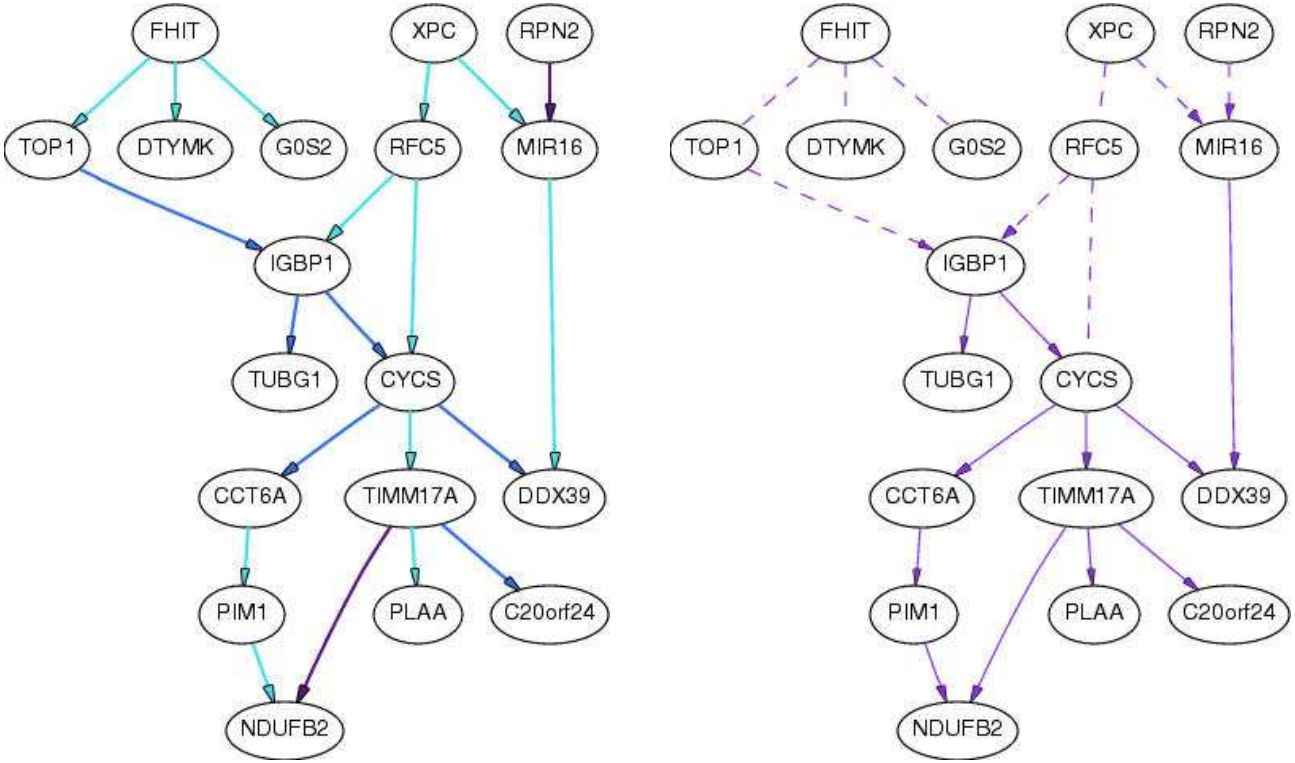
Bayesian networks with GAW15 problem 1

- Data from Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG: **Genetic analysis of genome-wide variation in human gene expression.** Nature 2004, **430:** 743-74
- Baseline expression levels of 8793 genes in immortalised B cells from 194 individuals in 14 Centre d'Etude du Polymorphisme Humane (CEPH) Utah pedigrees. Previous analysis of the data showed linkage and association and evidence of substantial individual variations.
- In particular, correlation was examined on expression levels of 31 genes and 25 target genes corresponding to two master regulatory regions
- In this analysis, we apply Bayesian network analysis to gain further insight into these findings. We identify strong dependences and therefore provide additional insight into the underlying relationships between the genes involved. More generally, the approach is expected to be applicable for integrated analysis of genes on biological pathways
- If the expression level of a given gene is regulated by certain proteins then it should be a function of the active levels of these proteins. Due to biological variability and measurement errors, the function would be stochastic rather than deterministic.
- Expression levels of genes are proxies for the activity level of the proteins they encode, although there are numerous examples where activation or silencing of a regulator is carried out by post-transcriptional protein modifications

Methods

- Gene expression levels, treated as continuous variables, can be assumed to follow a multivariate normal distribution, and to be consistent with a Bayesian network with linear Gaussian conditional densities.
- The prior of this network is characterised by a prior network reflecting our belief in the joint distribution of the variables in question, and equivalent sample size (ESS) effectively behaving as if it was calculated from a "prior" data set of that size. For instance, without a priori knowledge of the regulatory network, the prior network could be one where all expression levels are independent in order to avoid explicitly biasing the learning procedure to a particular edge.
- The common approach to the learning procedure starts with a training set and evaluates networks according to an asymptotically consistent scoring function that is obtained through the Bayesian framework.
- In the case of B-course software (http://b-course.hiit.fi) to be used here, discretisation of continuous data has been applied to capture the nonlinear relationship between variables and the choice of prior is such that the resulting ESS prior distribution is close to Jeffrey's prior. The software infers causal relationship according to the statistical dependence under some additional assumptions concerning latent variables. Mathematical details, including the definition of Jeffrey's prior, are given elsewhere
- The so-called causal structure assumes that dependencies between variables are due to causal relationships between variables in the model.

Results

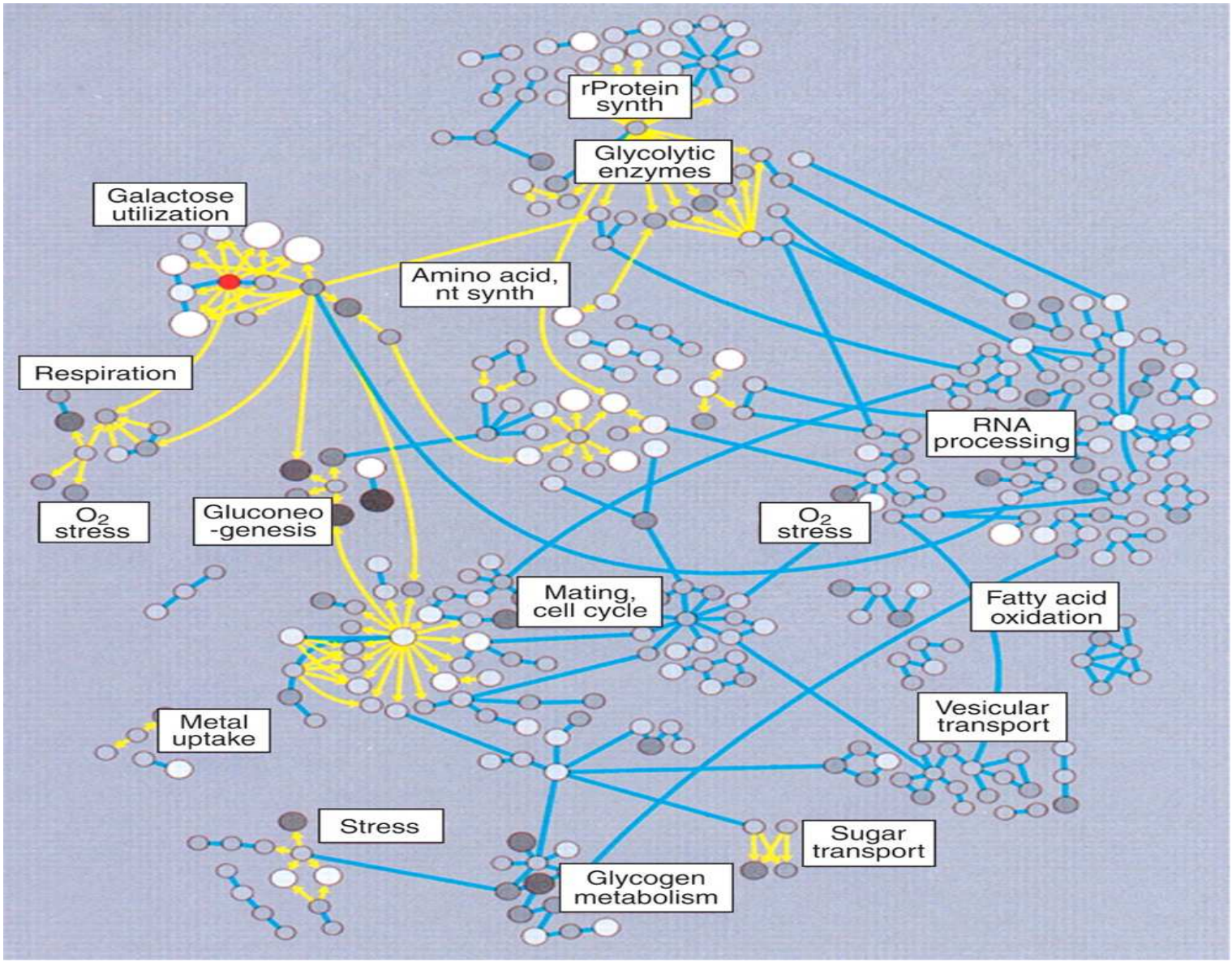


Left. Importance of the dependencies. **Right**. Solid arc has direct causal influence (direct meaning that causal influence is not mediated by any other variable that is included in the study). Dashed arc indicates there are two possibilities, but we do not know which holds. Dashed line without any arrow heads indicates there is a dependency but we do not know the reciprocal dependence. (Zhao et al. BMC Proc 1:S52, 2007, with b-course)

Summary

- The series of papers on these data stress the importance of Intermediate phenotypes. Without a priori biological hypothesis, it serves as an exploratory tool for subsequent confirmatory analysis.
- This particular analysis highlights the potential usefulness of pathway analysis but with great work to be done
- An apparent limitation of this work, though not uncommon in gene-expression studies, is the relatively small sample size used. To fully elucidate the biological pathways involved may be difficult, as for instance CYCS is involved in six pathways according to http://escience.invitrogen.com/ipath/.
- Statistical robustness and biological interpretability remain as the two main challenges for Bayesian network analyses, to which replication, bootstrap and benchmarking have been proposed.
- Our inference of gene networks also exploits the covariance structure of the data, like structural equation modelling, but is exploratory or hypothesis-generating rather than confirmatory or hypothesis-driven. A number of other software systems are of interest, e.g. ASIAN (a web-based regulatory network framework, http://eureka.cbrc.jp), deal.

Network Perturbation Model of Galactose Utilization in Yeast



Hood et al. (2004) Science

Metabolic pathways

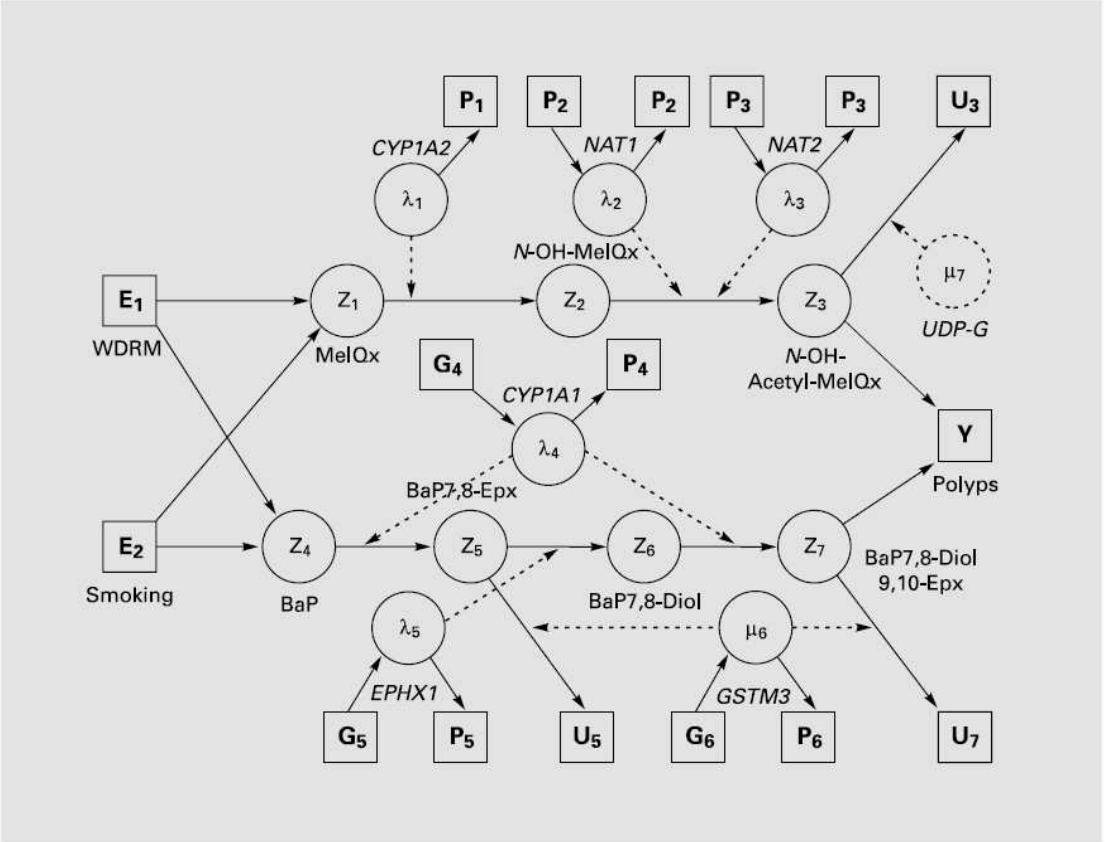


Fig. 1. Representation of the proposed metabolic pathways as a Directed Acyclic Graph (DAG). Boxes represent measured quantities for exposures (E), genotypes (G), metabolic phenotypes (P), the outcome polyps (Y), and urinary metabolites U (not included in this analysis). Circles represent the unobserved quantities, intermediate metabolites (Z) and metabolic activation (λ) and detoxification (μ) rates. Not shown are the genotypespecific population mean metabolic rates ($\overline{\lambda}s, \overline{\mu}s$), interpersonal variance in these rates σ^2, and metabolic phenotype measurement error variance (ω^2); see [9] for a discussion of these elements.

Conti et al. 2004. Hum Hered. With WinBUGS. The model was linked with GAW12 data generation. This example is generic in that a hypothesis-based pathway model can be fitted by elementary assumptions.

Bayesian networks

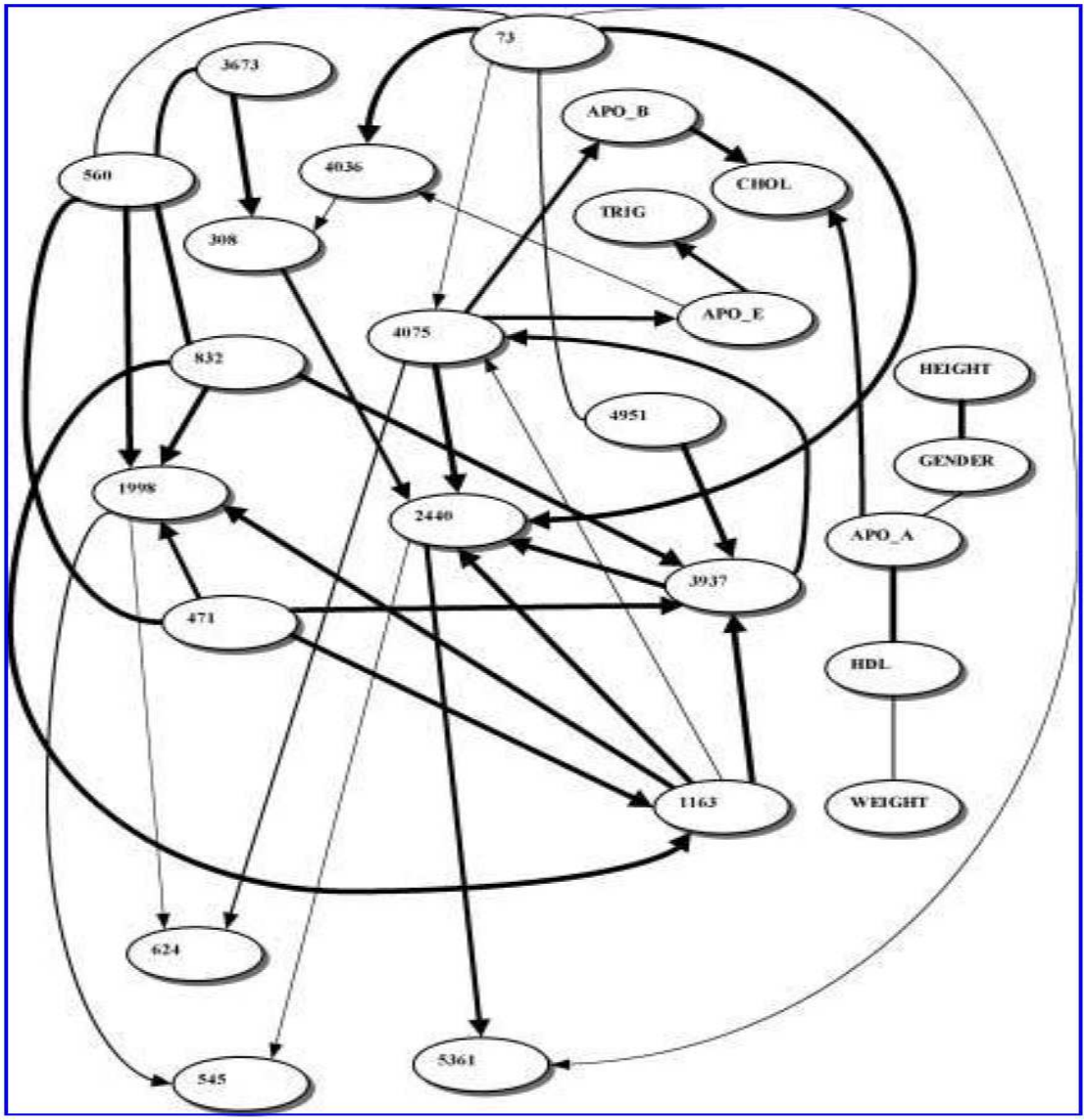


FIG. 1. Learned Belief network relating *APOE* SNPs to plasma **apoE** levels in Jackson, MS. Node legends: numbers refer to corresponding SNPs (see Fig. 1 in Nickerson *et al.* [2000] for an *APOE* SNP map). APO_E, APO_A, APO_B, TRIG, CHOL, and HDL stand for levels of apolipoproteins E, AI and B, triglycerides, cholesterol and high-density lipoprotein cholesterol, respectively. Line thickness corresponds to the relative edge strength (see Table 1.)

From Rodin et al. 2005, J Comp Biol. Exploratory Bayesian network analysis which involves both phenotype and genotype data and covariates

Health selection in civil servants

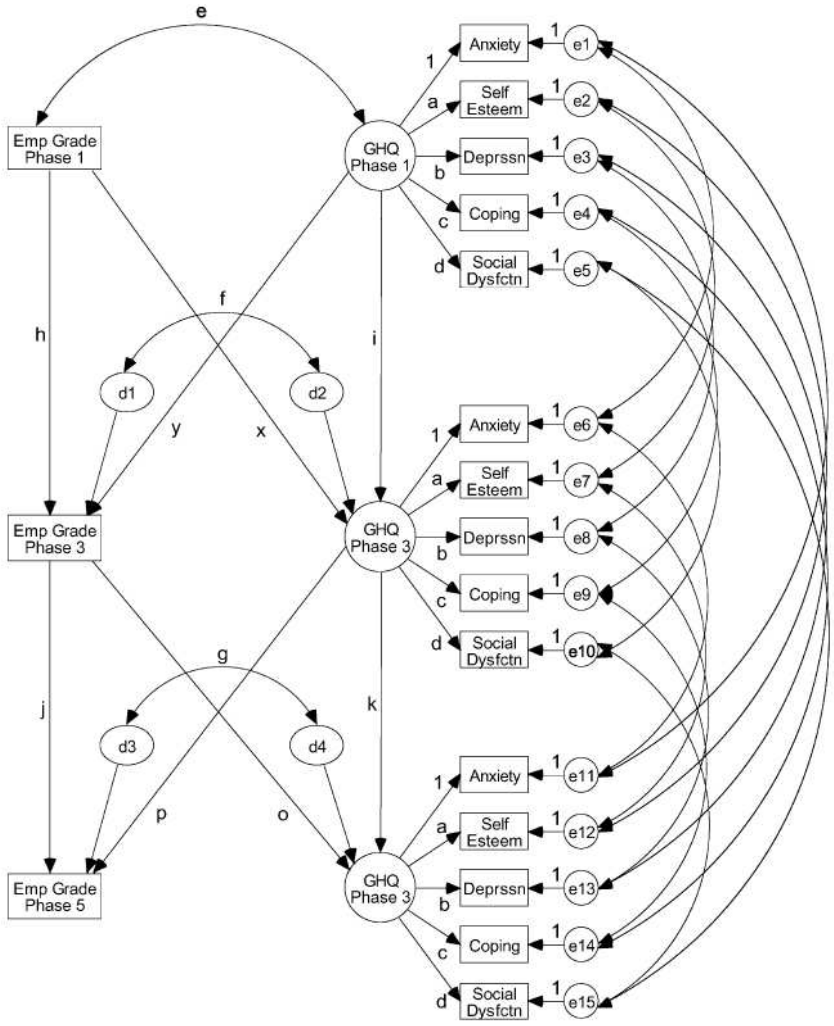


Fig. 1. Cross-lagged longitudinal analysis using structural equation models employment grade and GHQ at Whitehall phases 1, 3 and 5.

From Chandola et al. 2003, Soc Sci Med, with AMOS. It showed no selection. Similar findings were made with addition of phase 7 data.

Structural equation modelling

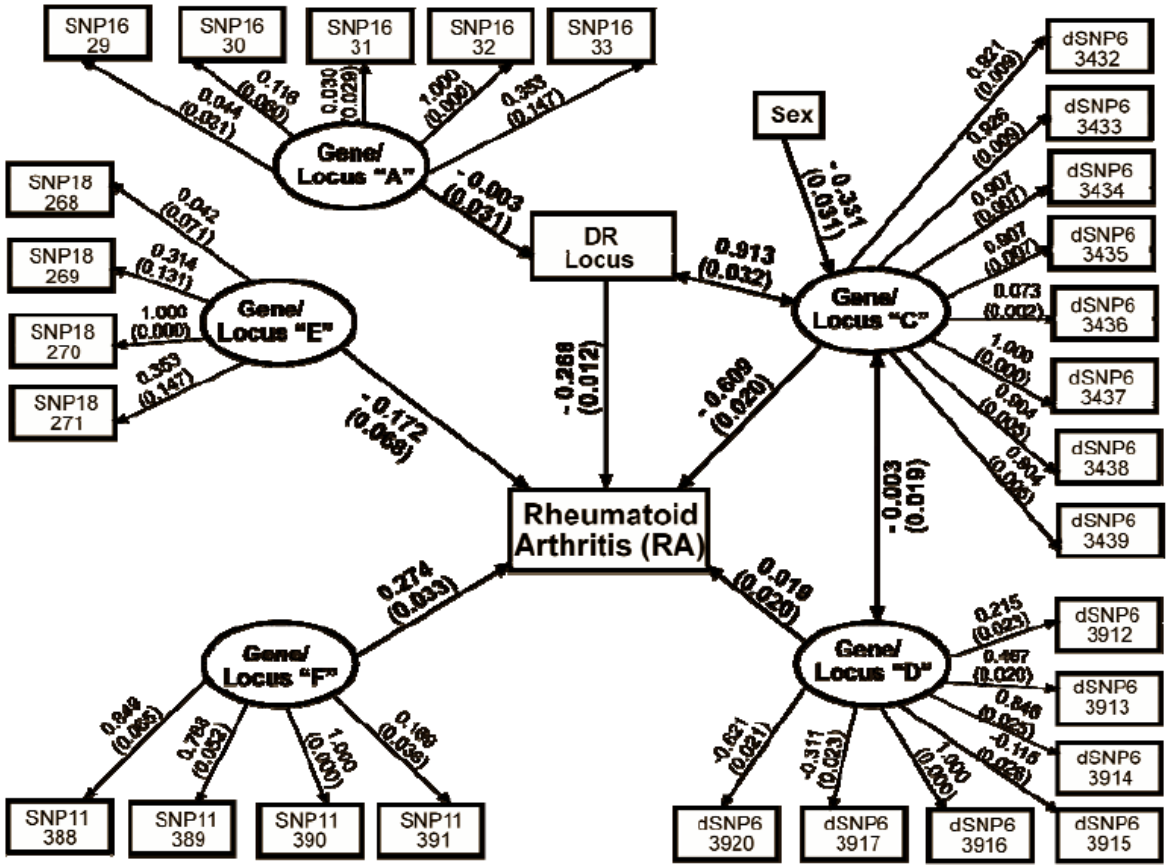
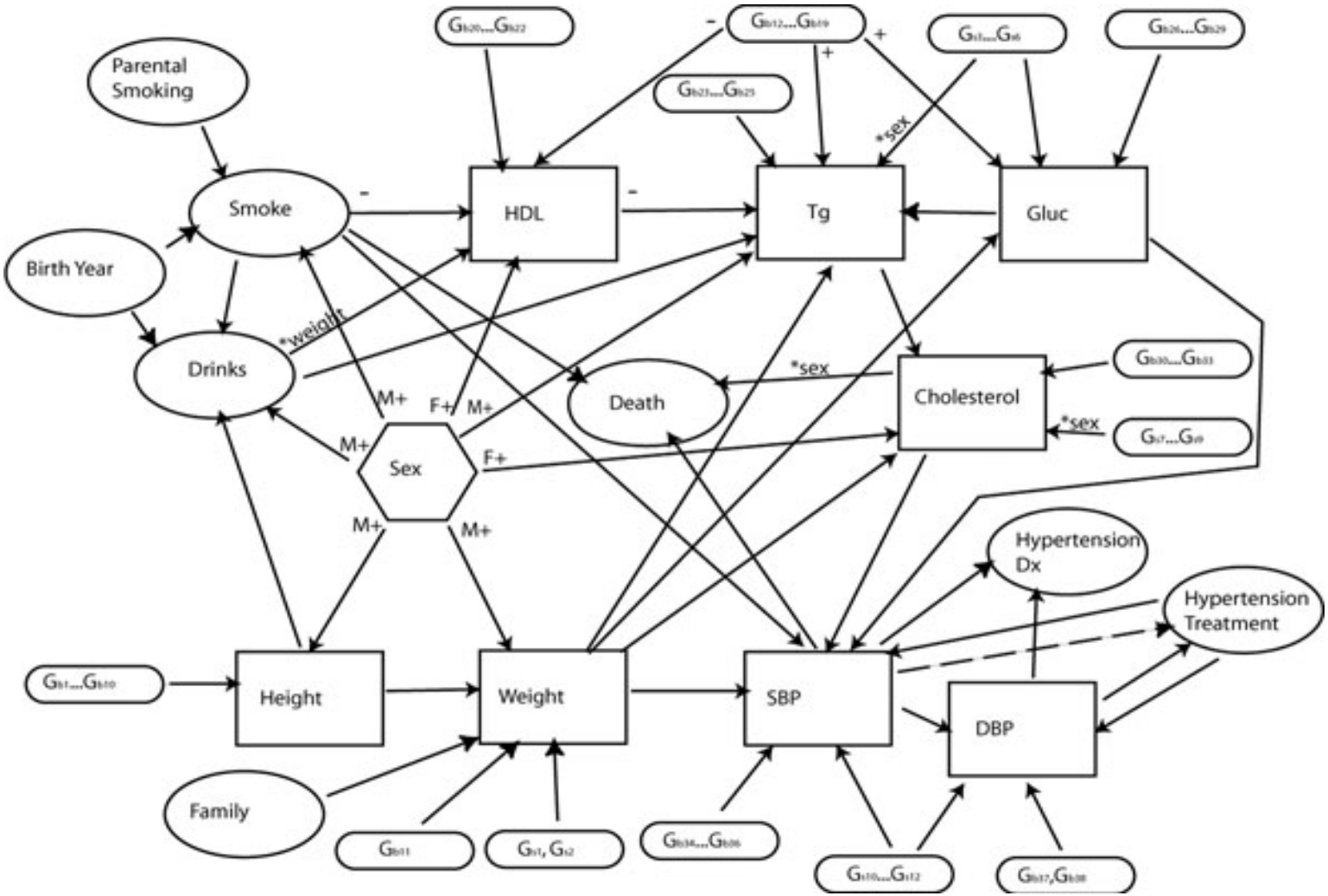


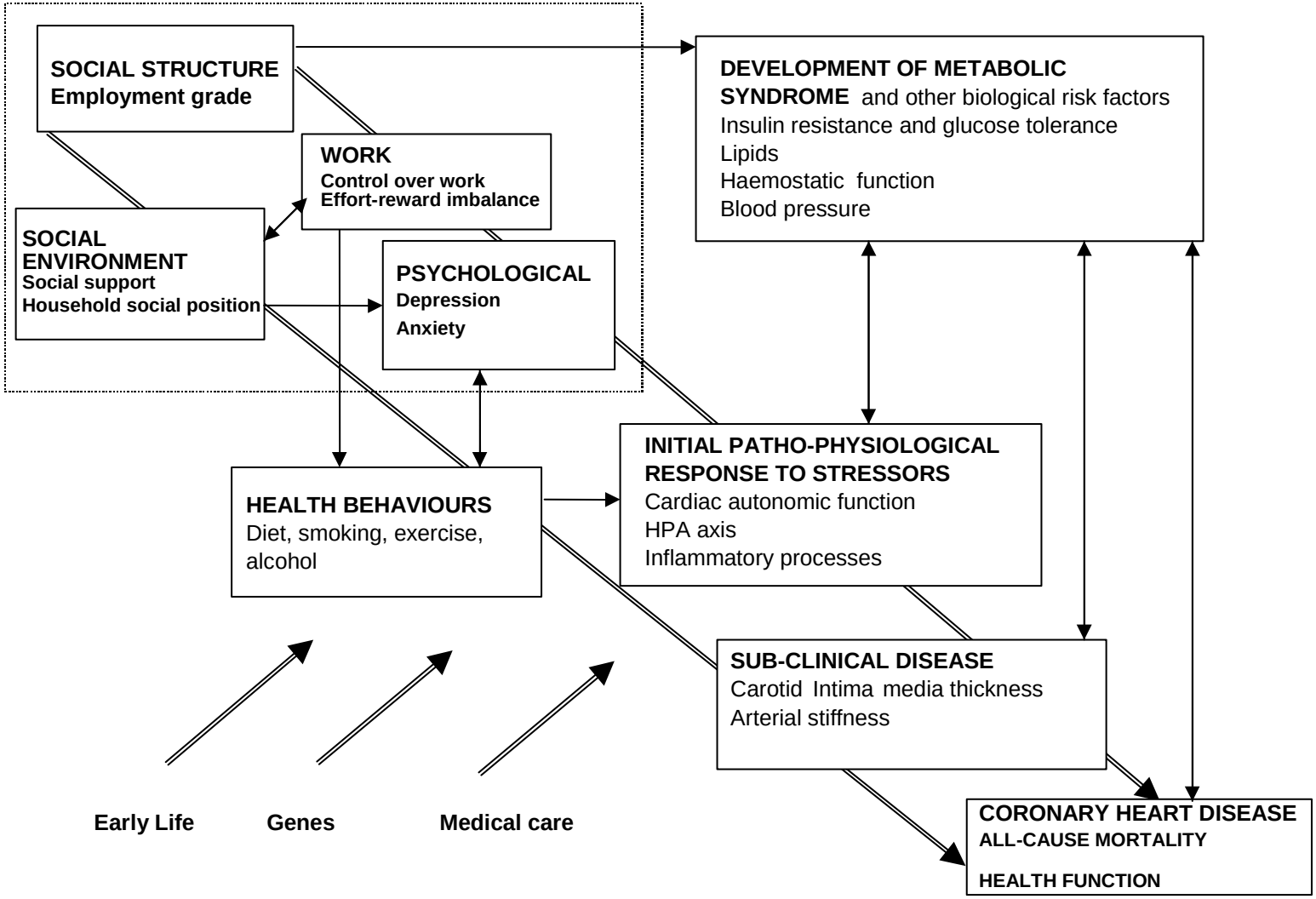
Figure 1. Evaluation of Rheumatoid Arthritis (RA) Data (Replicate #64) Using Structural Equation Modeling (SEM). Relationships in the measurement model between observed indicators (rectangles) and latent variables (ovals) are depicted by loading coefficients and standard errors in ( ) above single-headed unbolded arrows. Relationships in the structural model between latent variables and between latent and observed variables and the outcome, RA, are depicted by path coefficients and their standard errors in ( ) above single-headed bold arrows. Correlations are depicted by double-headed bold arrows.

From GAW15, with Mplus



GAW 13. Relationships between simulated traits and genes. Arrows indicate causal relationships between traits. Most correlations are positive, but a "-" indicates a negative correlation. An "*" and trait name next to an arrow indicates that the relationship is mediated by the named trait. Daw *et al. BMC Genetics* 2003 **4**(Suppl 1):S3

Whitehall II Study of 10,308 British Civil Servants from 1985



Master plan of research activities in the study